Методи чисельної оптимізації

Лекція 6: Адаптивні методи оптимізації

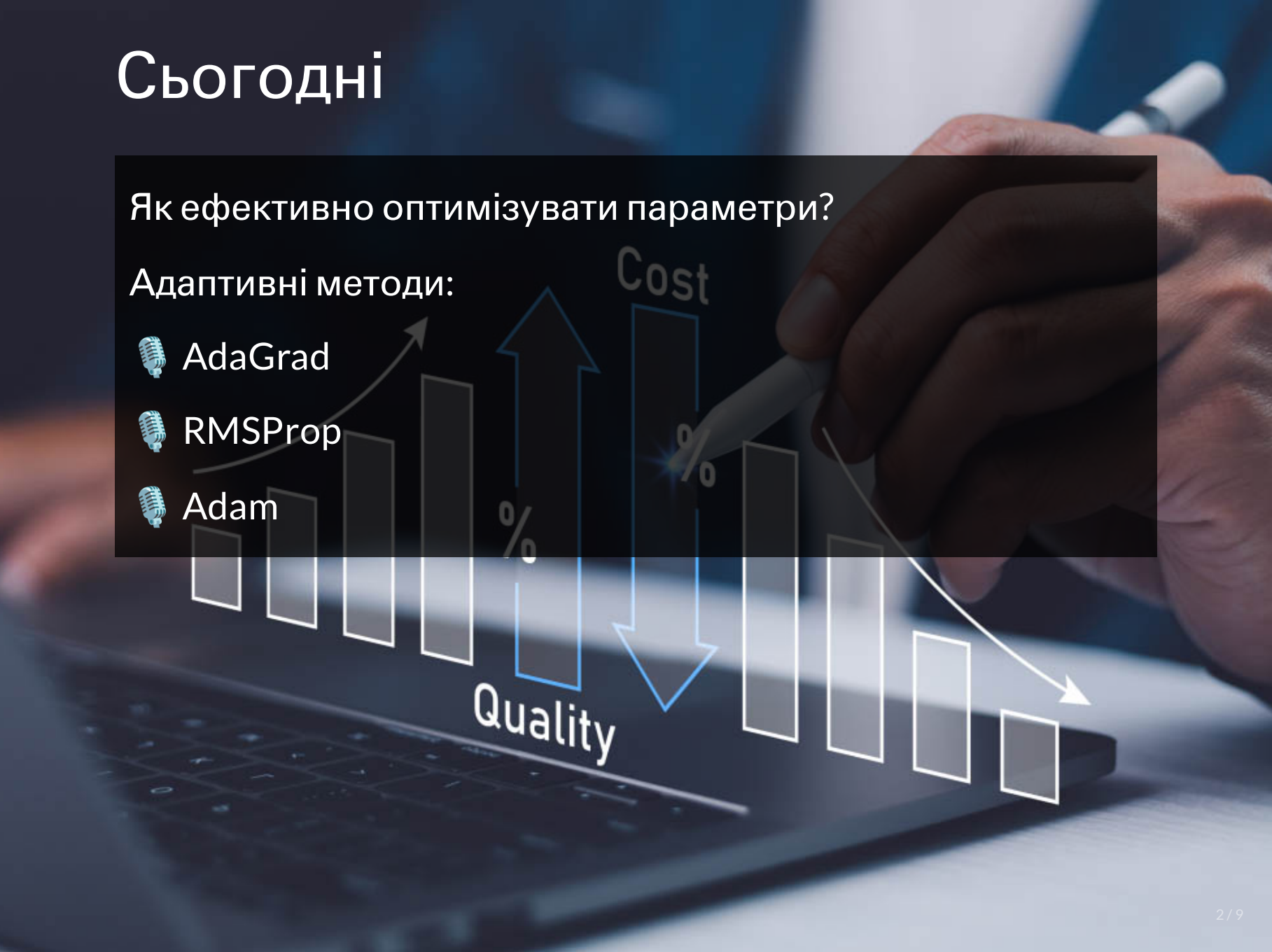
Кочура Юрій Петрович
iuriy.kochura@gmail.com
@y_kochura

Сьогодні

Як ефективно оптимізувати параметри?

Адаптивні методи:

- AdaGrad
- RMSProp
- Adam



Адаптивні методи

Адаптивні методи

Величина градієнтів часто сильно відрізняється між шарами нейронної мережі, тому глобальна швидкість навчання може не працювати належним чином.

Загальна ідея. Замість того, щоб використовувати однакову швидкість навчання для кожної ваги в нашій мережі, підтримуйте оцінку кращої швидкості окремо для кожної ваги.

Точний спосіб адаптації до швидкості навчання залежить від алгоритму, але більшість методів або пристосовуються до дисперсії ваг, або до локальної кривизни проблеми.

—
Credits: Aaron Defazio, Facebook AI Research

AdaGrad

Зменшення масштабу для кожного параметра на квадратний корінь із суми квадратів усіх його історичних значень.

$$r_t = r_{t-1} + g_t \odot g_t$$
$$W_{t+1} = W_t - \frac{\alpha}{\varepsilon + \sqrt{r_t}} \odot g_t$$

- AdaGrad позбавляє від необхідності вручну налаштовувати швидкість навчання. Більшість реалізацій використовують $\alpha = 0.01$ за замовчуванням.
- Добре, коли цільова функція опукла.
- r_t необмежено зростає під час навчання, що може спричинити зменшення розміру кроку та зрештою стати нескінченно малим.
- ε адитивна константа, яка гарантує, що ми не ділимо на 0.

RMSProp

Те саме, що AdaGrad, але накопичує експоненціально спадне середнє значення градієнта.

Ключова ідея. Нормалізувати за середньоквадратичним значенням градієнта

$$r_t = \rho r_{t-1} + (1-\rho) g_t \odot g_t$$
$$W_{t+1} = W_t - \frac{\alpha}{\varepsilon + \sqrt{r_t}} \odot g_t$$

- Ефективніший, коли цільова функція не є опуклою.

Adam: RMSprop з імпульсом

“Adaptive Moment Estimation”

Подібно до RMSProp з імпульсом, але з умовами корекції зсуву для першого та другого моментів.

$$
\begin{aligned}
p_t &= \rho_1 p_{t-1} + (1-\rho_1) g_t \\
\hat{p}_t &= \frac{p_t}{1-\rho_1^t} \\
r_t &= \rho_2 r_{t-1} + (1-\rho_2) g_t \odot g_t \\
\hat{r}_t &= \frac{r_t}{1-\rho_2^t} \\
W_{t+1} &= W_t - \alpha \frac{\hat{p}_t}{\varepsilon + \sqrt{\hat{r}_t}}
\end{aligned}
$$

- Хороші значення за замовчуванням $\rho_1 = 0.9$ і $\rho_2 = 0.999$.
- Подібно до того, як імпульс покращує SGD, він також покращує RMSProp.

Практична сторона

Для погано обумовлених задач Adam часто набагато кращий, ніж SGD.

Використовуйте Adam замість RMSprop завдяки очевидним перевагам імпульсу.

Але**, Adam** погано вивчений теоретично та має відомі недоліки**:**

- Зовсім не сходиться на деяких простих прикладах задач!
- Дає гіршу помилку узагальнення для багатьох задач комп’ютерного зору (наприклад, ImageNet)
- Потрібно більше пам'яті, ніж SGD
- Має 2 параметри моментів, тому може знадобитися додаткове налаштування

—
Credits: Aaron Defazio, Facebook AI Research

Демо

demo - losslandscape